                   Constrained Route Distribution for
    Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS)
          Internet Protcol (IP) Virtual Private Networks (VPNs)

Status of This Memo

   This document specifies an Internet standards track protocol for the
   Internet community, and requests discussion and suggestions for
   improvements. Please refer to the current edition of the "Internet
   Official Protocol Standards" (STD 1) for the standardization state
   and status of this protocol. Distribution of this memo is unlimited.

Copyright Notice

Abstract

   This document defines Multi-Protocol BGP (MP-BGP) procedures that
   allow BGP speakers to exchange Route Target reachability information.
   This information can be used to build a route distribution graph in
   order to limit the propagation of Virtual Private Network (VPN)
   Network Layer Reachability Information (NLRI) between different
   autonomous systems or distinct clusters of the same autonomous
   system. This document updates RFC 4364.

Table of Contents

1. Introduction

   In BGP/MPLS IP VPNs, PE routers use Route Target (RT) extended
   communities to control the distribution of routes into VRFs. Within
   a given iBGP mesh, PE routers need only hold routes marked with Route
   Targets pertaining to VRFs that have local CE attachments.

   It is common, however, for an autonomous system to use route
   reflection [2] in order to simplify the process of bringing up a new
   PE router in the network and to limit the size of the iBGP peering
   mesh.

   In such a scenario, as well as when VPNs may have members in more
   than one autonomous system, the number of routes carried by the
   inter-cluster or inter-as distribution routers is an important
   consideration.

   In order to limit the VPN routing information that is maintained at a
   given route reflector, RFC 4364 [3] suggests, in Section 4.3.3, the
   use of "Cooperative Route Filtering" [7] between route reflectors.
   This document extends the RFC 4364 [3] Outbound Route Filtering (ORF)
   work to include support for multiple autonomous systems and
   asymmetric VPN topologies such as hub-and-spoke.

   Although it would be possible to extend the encoding currently
   defined for the extended-community ORF in order to achieve this
   purpose, BGP itself already has all the necessary machinery for
   dissemination of arbitrary information in a loop-free fashion, both
   within a single autonomous system, as well as across multiple
   autonomous systems.

This document builds on the model described in RFC 4364 [3] and on
the concept of cooperative route filtering by adding the ability to
propagate Route Target membership information between iBGP meshes.
It is designed to supersede "cooperative route filtering" for VPN
related applications.

By using MP-BGP UPDATE messages to propagate Route Target membership
information, it is possible to reuse all of this machinery, including
route reflection, confederations, and inter-as information loop
detection.

Received Route Target membership information can then be used to
restrict advertisement of VPN NLRI to peers that have advertised
their respective Route Targets, effectively building a route
distribution graph. In this model, VPN NLRI routing information
flows in the inverse direction of Route Target membership
information.

This mechanism is applicable to any BGP NLRI that controls the
distribution of routing information by using Route Targets, such as
VPLS [9].

Throughout this document, the term NLRI, which expands to "Network
Layer Reachability Information", is used to describe routing
information carried via MP-BGP updates without any assumption of
semantics.

An NLRI consisting of {origin-as#, route-target} will be referred to
as RT membership information for the purpose of the explanation in
this document.

1.1. Terminology

This document uses a number of terms and acronyms specific to
Provider-Provisioned VPNs, including those specific to L2VPNs, L3VPNs
and BGP. Definitions for many of these terms may be found in the VPN
terminology document [10]. This section also includes some brief
acronym expansion and terminology to aid the reader.

AFI   Address Family Identifier (a BGP address type)

BGP   Border Gateway Protocol

BGP/MPLS VPN   A Layer 3 VPN implementation based upon BGP and MPLS

CE   Customer Edge (router)

    iBGP   Internal BGP (i.e., a BGP peering session that
                   connects two routers within an autonomous system)

    L2VPN   Layer 2 Virtual Private Network

    L3VPN   Layer 3 Virtual Private Network

    MP-BGP   MultiProtocol-Border Gateway Protocol

    MPLS   MultiProtocol Label Switching

    NLRI   Network Layer Reachability Information

    ORF   Outbound Route Filtering

    PE   Provider Edge (router)

    RT   Route Target (i.e., a BGP extended community that
                   conditions network layer reachability information with
                   VPN membership)

    SAFI   Subsequence Address Family Identifier (a BGP address
                   sub-type)

    VPLS   Virtual Private LAN Service

    VPN   Virtual Private Network

2. Specification of Requirements

   The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT",
   "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this
   document are to be interpreted as described in RFC 2119 [1].

3. NLRI Distribution

3.1. Inter-AS VPN Route Distribution

   In order to better understand the problem at hand, it is helpful to
   divide it in to its inter-Autonomous System (AS) and intra-AS
   components. Figure 1 represents an arbitrary graph of autonomous
   systems (a through j) interconnected in an ad hoc fashion. The
   following discussion ignores the complexity of intra-AS route
   distribution.

```
+-------------------------------+
| +---+    +---+    +---+        |
| | a | -- | b | -- | c |        |
| +---+    +---+    +---+        |
|   |        |                  |
|   |        |                  |
| +---+    +---+    +---+    +---+ |
| | d | -- | e | -- | f | -- | j | |
| +---+    +---+    +---+    +---+ |
|        /           |          |
|       /            |          |
| +---+    +---+    +---+        |
| | g | -- | h | -- | i |        |
| +---+    +---+    +---+        |
+-------------------------------+
```
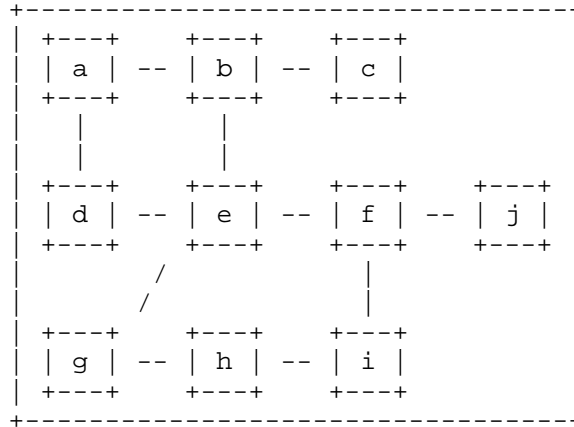
            Figure 1. Topology of autonomous systems

Let's consider the simple case of a VPN with CE attachments in ASes a
and i that uses a single Route Target to control VPN route
distribution. Ideally we would like to build a flooding graph for
the respective VPN routes that would not include nodes (c, g, h, j).
Nodes (c, j) are leafs ASes that do not require this information,
whereas nodes (g, h) are not in the shortest inter-as path between
(e) and (i) and thus should be excluded via standard BGP path
selection.

In order to achieve this, we will rely on ASa and ASi, generating a
NLRI consisting of {origin-as#, route-target} (RT membership
information). Receipt of such an advertisement by one of the ASes in
the network will signal the need to distribute VPN routes containing
this Route Target community to the peer that advertised this route.

Using RT membership information that includes both route-target and
originator AS number allows BGP speakers to use standard path
selection rules concerning as-path length (and other policy
mechanisms) to prune duplicate paths in the RT membership information
flooding graph, while maintaining the information required to reach
all autonomous systems advertising the Route Target.

In the example above, AS e needs to maintain a path to AS a in order
to flood VPN routing information originating from AS i and vice-
versa. It should, however, as default policy, prune less preferred
paths such as the longer path to ASi with as-path (g h i).

Extending the example above to include AS j as a member of the VPN
distribution graph would cause AS f to advertise 2 RT Membership
NLRIs to AS e, one containing origin AS i and one containing origin
AS j. Although advertising a single path would be sufficient to
guarantee that VPN information flows to all VPN member ASes, this is
not enough for the desired path selection choices. In the example
above, assume that (f j) is selected and advertised. Were that the
case, the information concerning the path (f i), which is necessary
to prune the arc (e g h i) from the route distribution graph, would
be missing.

As with other approaches for building distribution graphs, the
benefits of this mechanism are directly proportional to how "sparse"
the VPN membership is. Standard RFC2547 inter-AS behavior can be
seen as a dense-mode approach, to make the analogy with multicast
routing protocols.

3.2. Intra-AS VPN Route Distribution

As indicated above, the inter-AS VPN route distribution graph, for a
given route-target, is constructed by creating a directed arc on the
inverse direction of received Route Target membership UPDATEs
containing an NLRI of the form {origin-as#, route-target}.

Inside the BGP topology of a given autonomous-system, as far as
external RT membership information is concerned (route-targets where
the as# is not the local as), it is easy to see that standard BGP
route selection and advertisement rules [4] will allow a transit AS
to create the necessary flooding state.

Consider a IPv4 NLRI prefix, sourced by a single AS, which is
distributed via BGP within a given transit AS. BGP protocol rules
guarantee that a BGP speaker has a valid route that can be used for
forwarding of data packets for that destination prefix, in the
inverse path of received routing updates.

By the same token, and given that an {origin-as#, route-target} key
provides uniqueness between several ASes that may be sourcing this
route-target, BGP route selection and advertisement procedures
guarantee that a valid VPN route distribution path exists to the
origin of the Route Target membership information advertisement.

Route Target membership information that is originated within the
autonomous-system, however, requires more careful examination.
Several PE routers within a given autonomous-system may source the
same NLRI {origin-as#, route-target}, and thus default route
advertisement rules are no longer sufficient to guarantee that within
the given AS each node in the distribution graph has selected a
feasible path to each of the PEs that import the given route-target.

When processing RT membership NLRIs received from internal iBGP
peers, it is necessary to consider all available iBGP paths for a
given RT prefix, for building the outbound route filter, and not just
the best path.

In addition, when advertising Route Target membership information
sourced by the local autonomous system to an iBGP peer, a BGP speaker
shall modify its procedure to calculate the BGP attributes such that
the following apply:

i.   When advertising RT membership NLRI to a route-reflector client,
      the Originator attribute shall be set to the router-id of the
      advertiser, and the Next-hop attribute shall be set of the local
      address for that session.

ii.   When advertising an RT membership NLRI to a non-client peer, if
      the best path as selected by the path selection procedure
      described in Section 9.1 of the base BGP specification [4] is a
      route received from a non-client peer, and if there is an
      alternative path to the same destination from a client, the
      attributes of the client path are advertised to the peer.

The first of these route advertisement rules is designed such that
the originator of an RT membership NLRI does not drop an RT
membership NLRI that is reflected back to it, thus allowing the route
reflector to use this RT membership NLRI in order to signal the
client that it should distribute VPN routes with the specific target
towards the reflector.

The second rule allows any BGP speaker present in an iBGP mesh to
signal the interest of its route reflection clients in receiving VPN
routes for that target.

These procedures assume that the autonomous-system route reflection
topology is configured such that IPv4 unicast routing would work
correctly. For instance, route reflection clusters must be
contiguous.

An alternative solution to the procedure given above would have been
to source different routes per PE, such as NLRI of the form
{originator-id, route-target}, and to aggregate them at the edge of
the network. The solution adopted is considered advantageous over
the former in that it requires less routing-information within a
given AS.

4. Route Target Membership NLRI Advertisements

   Route Target membership NLRI is advertised in BGP UPDATE messages
   using the MP_REACH_NLRI and MP_UNREACH_NLRI attributes [5]. The
   [AFI, SAFI] value pair used to identify this NLRI is (AFI=1,
   SAFI=132).

   The Next Hop field of MP_REACH_NLRI attribute shall be interpreted as
   an IPv4 address whenever the length of NextHop address is 4 octets,
   and as a IPv6 address whenever the length of the NextHop address is
   16 octets.

   The NLRI field in the MP_REACH_NLRI and MP_UNREACH_NLRI is a prefix
   of 0 to 96 bits, encoded as defined in Section 4 of [5].

   This prefix is structured as follows:

```
        +-------------------------------+
        | origin as        (4 octets)   |
        +-------------------------------+
        | route target     (8 octets)   |
        +                               +
        |                               |
        +-------------------------------+
```

   Except for the default route target, which is encoded as a zero-
   length prefix, the minimum prefix length is 32 bits. As the origin-
   as field cannot be interpreted as a prefix.

   Route targets can then be expressed as prefixes, where, for instance,
   a prefix would encompass all route target extended communities
   assigned by a given Global Administrator [6].

   The default route target can be used to indicate to a peer the
   willingness to receive all VPN route advertisements such as, for
   instance, the case of a route reflector speaking to one of its PE
   router clients.

5. Capability Advertisement

   A BGP speaker that wishes to exchange Route Target membership
   information must use the Multiprotocol Extensions Capability Code, as
   defined in RFC 2858 [5], to advertise the corresponding (AFI, SAFI)
   pair.

   A BGP speaker MAY participate in the distribution of Route Target
   information without using the learned information for purposes of VPN
   NLRI output route filtering, although this is discouraged.

6. Operation

   A VPN NLRI route should be advertised to a peer that participates in
   the exchange of Route Target membership information if that peer has
   advertised either the default Route Target membership NLRI or a Route
   Target membership NLRI containing any of the targets contained in the
   extended communities attribute of the VPN route in question.

   When a BGP speaker receives a BGP UPDATE that advertises or withdraws
   a given Route Target membership NLRI, it should examine the RIB-OUTs
   of VPN NLRIs and re-evaluate the advertisement status of routes that
   match the Route Target in question.

   A BGP speaker should generate the minimum set of BGP VPN route
   updates (advertisements and/or withdrawls) necessary to transition
   between the previous and current state of the route distribution
   graph that is derived from Route Target membership information.

   As a hint that initial RT membership exchange is complete,
   implementations SHOULD generate an End-of-RIB marker, as defined in
   [8], for the Route Target membership (afi, safi), regardless of
   whether graceful-restart is enabled on the BGP session. This allows
   the receiver to know when it has received the full contents of the
   peer's membership information. The exchange of VPN NLRI should
   follow the receipt of the End-of-RIB markers.

   If a BGP speaker chooses to delay the advertisement of BGP VPN route
   updates until it receives this End-of-RIB marker, it MUST limit that
   delay to an upper bound. By default, a 60 second value should be
   used.

7. Deployment Considerations

   This mechanism reduces the scaling requirements that are imposed on
   route reflectors by limiting the number of VPN routes and events that
   a reflector has to process to the VPN routes used by its direct
   clients. By default, a reflector must scale in terms of the total
   number of VPN routes present on the network.

   This also means that it is now possible to reduce the load imposed on
   a given reflector by dividing the PE routers present on its cluster
   into a new set of clusters. This is a localized configuration change
   that need not affect any system outside this cluster.

   The effectiveness of RT-based filtering depends on how sparse the VPN
   membership is.

   The same policy mechanisms applicable to other NLRIs are also
   applicable to RT membership information. This gives a network
   operator the option of controlling which VPN routes get advertised in
   an inter-domain border by filtering the acceptable RT membership
   advertisements inbound.

   For instance, in the inter-as case, it is likely that a given VPN is
   connected only to a subset of all participating ASes. The only
   current mechanism to limit the scope of VPN route flooding is through
   manual filtering on the external BGP border routers. With the
   current proposal, such filtering can be performed according to the
   dynamic Route Target membership information.

   In some inter-as deployments, not all RTs used for a given VPN have
   external significance. For example, a VPN can use a hub RT and a
   spoke RT internally to an autonomous-system. The spoke RT does not
   have meaning outside this AS, so it may be stripped at an external
   border router. The same policy rules that result in extended
   community filtering can be applied to RT membership information in
   order to avoid advertising an RT membership NLRI for the spoke-RT in
   the example above.

   Throughout this document, we assume that autonomous-systems agree on
   an RT assignment convention. RT translation at the external border
   router boundary is considered a local implementation decision, as it
   should not affect inter-operability.

8. Security Considerations

   This document does not alter the security properties of BGP-based
   VPNs. However, note that output route filters built from RT
   membership information NLRIs are not intended for security purposes.
   When exchanging routing information between separate administrative
   domains, it is a good practice to filter all incoming and outgoing
   NLRIs by some other means in addition to RT membership information.
   Implementations SHOULD also provide means to filter RT membership
   information.

9. Acknowledgements

   This proposal is based on the extended community route filtering
   mechanism defined in [7].

   Ahmed Guetari was instrumental in defining requirements for this
   proposal.

   The authors would also like to thank Yakov Rekhter, Dan Tappan, Dave
   Ward, John Scudder, and Jerry Ash for their comments and suggestions.

10. References

10.1. Normative References

   [1]   Bradner, S., "Key words for use in RFCs to Indicate Requirement
         Levels", BCP 14, RFC 2119, March 1997.

   [2]   Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An
         Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April
         2006.

   [3]   Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks
         (VPNs)", RFC 4364, February 2006.

   [4]   Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4
         (BGP-4)", RFC 4271, January 2006.

   [5]   Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol
         Extensions for BGP-4", RFC 2858, June 2000.

   [6]   Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended
         Communities Attribute", RFC 4360, February 2006.

10.2. Informative References

   [7] Chen, E. and Y. Rekhter, "Cooperative Route Filtering Capability
        for BGP-4", Work in Progress, December 2004.

   [8] Sangli, S., Rekhter, Y., Fernando, R., Scudder, J., and E. Chen,
        "Graceful Restart Mechanism for BGP", Work in Progress, June
        2004.

   [9] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service", Work
        in Progress, April 2005.

   [10] Andersson, L. and T. Madsen, "Provider Provisioned Virtual
        Private Network (VPN) Terminology", RFC 4026, March 2005.

Authors' Addresses

   Pedro Marques
   Juniper Networks
   1194 N. Mathilda Ave.
   Sunnyvale, CA 94089
   US

   EMail: roque@juniper.net

   Ronald Bonica
   Juniper Networks
   1194 N. Mathilda Ave.
   Sunnyvale, CA 94089
   US

   EMail: rbonica@juniper.net

   Luyuan Fang
   Cisco Systems, Inc.
   300 Beaver Brook Road
   Boxborough, MA 01719
   US

   EMail: lufang@cisco.com

   Luca Martini
   Cisco Systems, Inc.
   9155 East Nichols Avenue, Suite 400
   Englewood, CO 80112
   US

   EMail: lmartini@cisco.com

   Robert Raszuk
   Cisco Systems, Inc.
   170 West Tasman Dr
   San Jose, CA 95134
   US

   EMail: rraszuk@cisco.com

   Keyur Patel
   Cisco Systems, Inc.
   170 West Tasman Dr
   San Jose, CA 95134
   US

   EMail: keyupate@cisco.com

   Jim Guichard
   Cisco Systems, Inc.
   300 Beaver Brook Road
   Boxborough, MA 01719
   US

   EMail: jguichar@cisco.com

Intellectual Property

   The IETF takes no position regarding the validity or scope of any
   Intellectual Property Rights or other rights that might be claimed to
   pertain to the implementation or use of the technology described in
   this document or the extent to which any license under such rights
   might or might not be available; nor does it represent that it has
   made any independent effort to identify any such rights. Information
   on the procedures with respect to rights in RFC documents can be
   found in BCP 78 and BCP 79.

   Copies of IPR disclosures made to the IETF Secretariat and any
   assurances of licenses to be made available, or the result of an
   attempt made to obtain a general license or permission for the use of
   such proprietary rights by implementers or users of this
   specification can be obtained from the IETF on-line IPR repository at
   http://www.ietf.org/ipr.

   The IETF invites any interested party to bring to its attention any
   copyrights, patents or patent applications, or other proprietary
   rights that may cover technology that may be required to implement
   this standard. Please address the information to the IETF at
   ietf-ipr@ietf.org.